How does image noise affect actual and predicted human gaze allocation in assessing image quality?

Florian Röhrbein[1], Peter Goddard[2], Michael Schneider[1],Georgina James[2], Kun Guo[2]

[1]Institut für Informatik VI, Technische Universität München, Germany

[2]School of Psychology, University of Lincoln, UK

Corresponding Author:

Dr. Kun Guo

School of Psychology, University of Lincoln, Lincoln, LN6 7TS, UK

Email address: kguo@lincoln.ac.uk

Tel: +44-1522-886294

Fax: +44-1522-886026

Abstract

A central research question in natural vision is how to allocate fixation to extract informative cues for scene perception. With high quality images, psychological and computational studies have made significant progress to understand and predict human gaze allocation in scene exploration. However, given perceptual processing strategy changes with external noise level, we need to revisit this question with realistic visual signals often embedded in natural distortions. In this study, we applied Gaussian low-pass filter, circular averaging filter and Additive Gaussian white noise to systematically distort man-made and natural scenes, and recorded participants' gaze behaviour in assessing perceived image qualities. Our analysis showed that in comparison with original high quality images, distorted images attracted fewer numbers of fixations but longer fixation durations, shorter saccades and stronger central fixation bias. This systematically varied gaze pattern in scene viewing was mainly determined by noise intensity, and the same noise type could have different impact on the perceived image quality and associated gaze pattern to different scene categories. We furthered compared four high performing visual attention models in predicting human gaze allocation in degraded scenes, and found that model performance lacked human-like sensitivity to noise type and intensity, and was considerably worse than human performance measured as inter-observer variance. Furthermore, the central fixation bias is a major predictor for human gaze allocation, which becomes more prominent with increased noise intensity. Our results indicate a crucial role of external noise intensity in determining scene-viewing gaze behaviour, which should be considered in the development of realistic human-vision-inspired attention models.

Keywords: Natural scene; image distortion; image quality; gaze behaviour; visual attention model

Introduction

When exploring natural surroundings, we do not direct our attention evenly or randomly to different parts of the scene. Instead, we make a series of saccades to direct a limited number of fixations to local regions that are informative or interesting to us. The preferred regions within a scene are often inspected earlier and attract more fixations (Henderson, 2007). Such gaze allocation provides a real-time behaviour index of on-going perceptual and cognitive processing and is a sensitive index of our attention, motivation, and preference, especially when exploring scenes of high ecological validity (Isaacowitz, 2006; Henderson, 2007). One central research question in this active visual exploration process is to understand how we choose the fixated local regions in the scene.

Many empirical studies have suggested that both bottom-up local saliency computation and top-down cognitive processes are actively involved in determining our fixations in scene exploration. Specifically, the choice of foveated local region is heavily influenced by local low-level image saliency (e.g., local image colour, intensity, contrast, spatial frequency, and structure). We tend to avoid low-contrast and homogeneous 'predictable' regions in natural scenes, and bias our fixation to local features with high-contrast, high spatial frequency, high edge density, and complex local structure (e.g., curved lines, edges and corners, as well as occlusions or isolated spots) (Mannan, Ruddock, & Wooding, 1995, 1996; Reinagel & Zador 1999; Parkhurst & Niebur 2003; Krieger et al., 2000; Acik et al., 2009), or to local regions deviated from surrounding image statistics (Einhäuser et al., 2006). On the other hand, top-down factors, such as expectation, memory, semantic and task-related knowledge, could significantly modulate gaze allocation in scene exploration (Henderson, 2007; Tatler et al., 2011; Guo et al., 2012; Pollux, Hall, & Guo, 2014).

These experimental findings complement the development of computational models for predicting where people look in natural vision. Closely resembling our knowledge about neural processing in early visual system, the widely cited bottom-up saliency model (Itti & Koch, 2000) compares local image intensity, colour and orientation through centre-surround filtering at eight spatial scales, combines them into a single salience (conspicuity) map with a winner-take-all network and inhibition-of-return, and then produces a sequence of predicted fixations that scan the scene in order of decreasing salience. To improve its relatively low level of predictive power (e.g., 57% – 68% correct fixation prediction in some scene free-viewing tasks,

Betz et al., 2010), some top-down processing such as scene context (contextual guidance model, Torralba et al., 2006; context-aware saliency, Goferman, Zelnik-Manor, & Tal, 2012), object detection (Judd et al., 2009) and natural statistics (Kanan et al., 2009) are later incorporated into the model. Incorporating these top-down cues does not necessarily sacrifice the computational precision of the original saliency map model, or even alter the basic structure of the approach (Navalpakkam & Itti, 2005). Specifically, combining both bottom-up saliency-driven information and top-down natural scene understanding would greatly improve gaze predictions in a real-word image search task (Kanan et al., 2009). It seems that humans utilize both local image saliency and global scene understanding in guiding eye movements to efficiently sample scene information.

These experimental findings and computational models of visual attention in scene perception are derived mainly from studies using high-quality images in laboratory settings. Real-world scene perception, however, often involves selecting, extracting and processing diagnostic information from a noisy environment (e.g., due to bad weather condition). Typically, the images and videos we view daily are subject to a variety of distortions during acquisition, compression, storage, transmission and reproduction, any of which will degrade visual quality. It is proposed that most distortion processes would disturb natural image statistics (Sheikh, Bovik, & de Veciana, 2005) and may attract attention away from local regions that are salient in undistorted images. Furthermore, our perceptual processing strategy tends to change with the level of external noise, independent of the observer's internal noise (Allard & Cavanagh, 2012).

Considering that our visual system has evolved and/or learned over time to process visual signals embedded in natural distortions, it is reasonable to assume that we should have developed a near-optimal processing strategy for visual signals corrupted by these distortions. So far only a handful of psychophysical and computational studies have attempted to investigate our perceptual sensitivity to image blur (e.g., Watson & Ahumada, 2011) and image resolution (e.g., Castelhano & Henderson, 2008; Torralba, 2009). These studies have shown that we could essentially classify natural scenes or understand scene gist at a very low resolution (up to 16×16 pixels depending on image complexity), suggesting that we might use the same diagnostic visual cues in low- and high-resolution scenes. One recent eye-tracking study further showed that although low-resolution images attracted fewer

fixations with shorter saccade length, the location of fixations on low-resolution images tended to be similar to and predictive of fixations on high-resolution images (Judd, Durand, & Torralba, 2011). On the other hand, some studies have observed that viewing of noisy images (e.g., applying masking, low- or high-pass spatial frequency filters to different image regions) was associated with shorter saccade amplitudes and longer fixation durations (Mannan, Ruddock, & Wooding, 1995; Pomplun, Reingold, & Shen, 2001; Loschly & McConkie, 2002; van Diepen & d'Ydewalle, 2003; Nuthmann, 2013), indicating human fixation distribution in image viewing may change with image noise.

These findings are potentially very significant to refine models of visual attention in scene perception. However, the generalisation of them is limited by methodological issues such as use of a narrow range of scenes (different categories of natural scenes have different scene statistics which may be subject to different impact by the same distortion type, e.g., the appearance of high spatial frequency stimuli is more affected by blur than low spatial frequency stimuli), and concentration on the manipulation of image parameters (e.g., resolution) rather than perceptually perceived image quality. It is unclear how different types and levels of image distortion would impact on perceived image quality, gaze pattern used to assess image quality, and predictive power of visual attention models. As we always assume that our brain has evolved to efficiently code and transmit information from natural surroundings, to determine what would be an efficient code in natural vision, it is essential to know how variance in image noise would affect scene saliency computation, and cognitive processes involved in sampling and encoding degraded scene information. Such research also meets strong and present interest in computer vision and signal processing to develop human-vision-inspired foveated active artificial vision systems and image/video quality assessment algorithms (e.g., Winkler, 2012) that will benefit numerous applications, such as enhancing the multimedia experience of human consumers and improving the efficiency of surveillance systems.

In this study we combined psychophysical, high-speed eye-tracking and computational approaches to investigate how different image distortions affected our gaze behaviour in assessing the perceived image qualities and the predictive power of computational saliency models. In the eye-tracking experiment, we applied a Gaussian low-pass filter, circular averaging filter and additive Gaussian white noise to systematically distort both man-made and natural landscape scenes, and recorded

participants' gaze patterns in evaluating the perceived quality of the distorted images. In the following computational experiment, we applied various state-of-the-art computational models of visual attention, such as Judd model (Judd et al., 2009), Erdem model (Erdem & Erdem, 2013), Graph-based visual saliency model (Harel, Koch, & Perona, 2007) and Adaptive whitening saliency model (Garcia-Diaz et al., 2012a, 2012b), to these natural images of varying distortion, and systematically compared their performance in predicting human gaze allocation in viewing of degraded images.

Experiment 1: Eye-tracking study

Methods

Twenty-four undergraduate students (16 female, 8 male), age ranging from 18 to 25 years old with the mean of 20.67±2.48 (Mean±SEM), volunteered to participate in this study. All participants had normal or corrected-to-normal visual acuity, and normal colour vision (checked with Ishihara's Tests for Colour Deficiency, 24 Plates Edition). The Ethical Committee in School of Psychology, University of Lincoln approved this study. Written informed consent was obtained from each participant, and all procedures complied with the British Psychological Society Code of Ethics and Conduct and with the World Medical Association Helsinki Declaration as revised in October 2008.

Digitized colour scene images were presented through a ViSaGe graphics system (Cambridge Research Systems, UK) and displayed on a non-interlaced gamma-corrected colour monitor (30 cd/m^2 background luminance, 100 Hz frame rate, Mitsubishi Diamond Pro 2070SB) with the resolution of 1024 × 768 pixels. At a viewing distance of 57 cm, the monitor subtended a visual angle of 40 × 30 deg.

10 man-made scenes and 10 natural landscape scenes were sampled from the author's collection based on the DynTex database (Péteri, Fazekas, & Huiskes, 2010) (Fig.1). The original high quality images had identical size of 768 × 576 pixels. To systematically degrade the perceived image quality, we manipulated each original image with three different types of distortion or noise (average noise, Gaussian blur, and additive Gaussian noise) to cover the most common variants, and each distortion type had two noise intensities (weak and strong noise). Specifically, average noise was created by applying a circular averaging filter with a radius of 2 for weak noise intensity (Avg W) and a radius of 10 for strong noise intensity (Avg S). Gaussian blur

was created with a rotationally symmetric Gaussian low-pass filter of size 20 with standard deviation of 2 for weak noise intensity (Gaussian W) and 8 for strong noise intensity (Gaussian S). Additive noise distortion was created by adding white Gaussian noise with a different signal-to-noise ratio to the original image, 10 dB for weak noise intensity (SNR W) and 0 dB for strong noise intensity (SNR S). These noise intensity levels for different distortion types were determined previously in a pilot study by asking an independent group of 10 participants to evaluate the impact of noise level on the perceived image quality on a 7-point Likert scale. For the chosen weak or strong noise intensity, the perceived image quality was comparable between images and different distortion types.

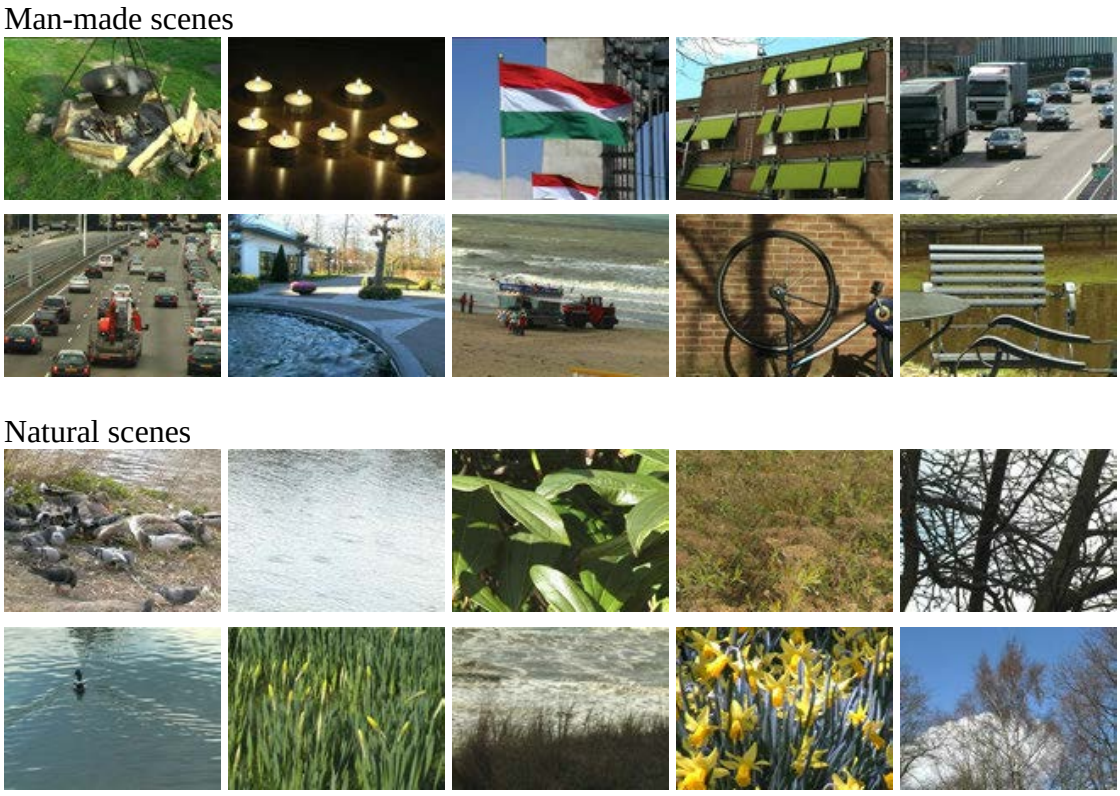

Figure 1. Original images of man-made and natural scenes used in this study.

As a result, for each of 20 original high-quality images a set of six degraded variants (3 noise type × 2 noise intensity) was created (see Fig. 2 for examples). In total 140 scene images were generated for the testing session. These images were gamma corrected and displayed once in a random order during the testing.

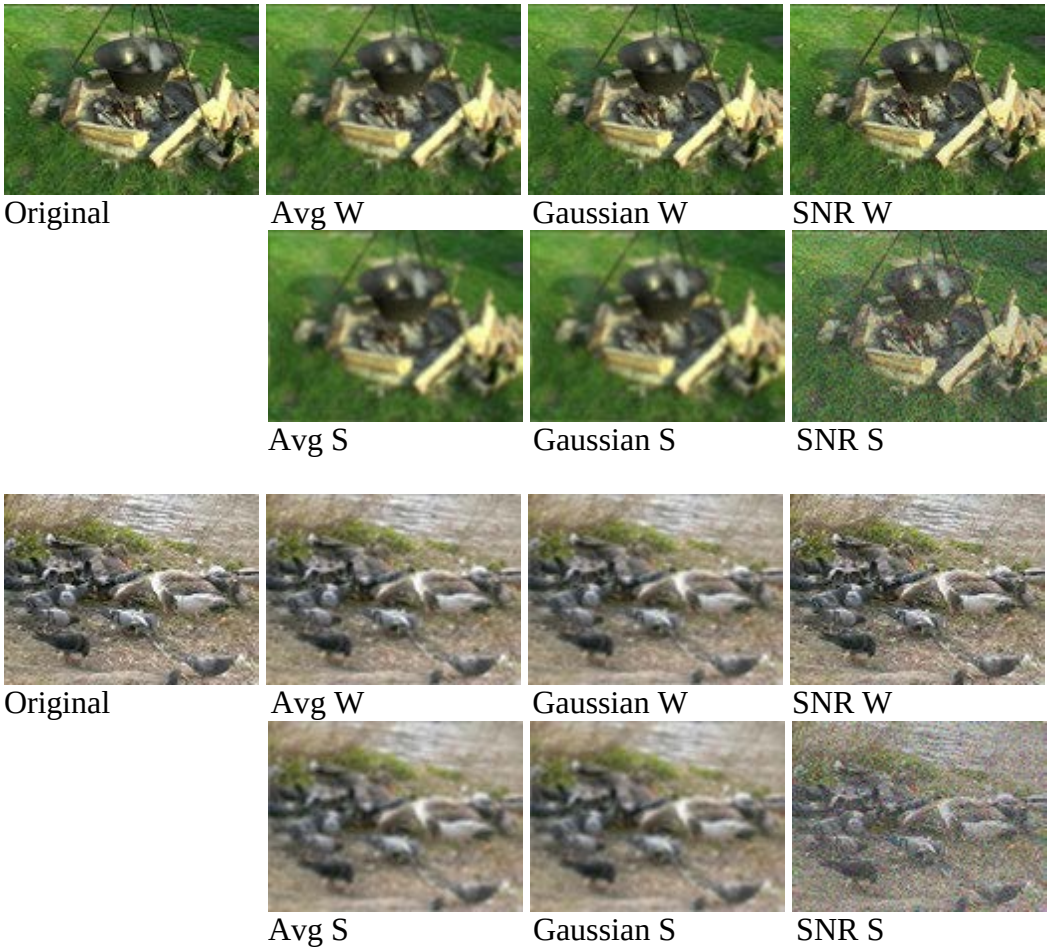

Figure 2. Examples of original scene images and their distorted variants (see text for detailed distortion types and noise intensities).

A self-paced task was used to mimic natural viewing condition. During the experiments the participants sat in a chair with their head restrained by a chin-rest, and viewed the display binocularly. Their horizontal and vertical eye positions from the self-reported dominant eye (determined through the Hole-in-the-Card test or the Dolman method if necessary) were measured using Cambridge Research Systems High-Speed Video Eyetracker Toolbox (250 Hz sampling frequency, 0.25 deg accuracy; Cambridge Research Systems, UK). To calibrate eye movement signals, a small red fixation point (FP, 0.3 deg diameter, 15 cd/m^2 luminance) was displayed randomly at one of 9 positions (3 × 3 matrix) across the monitor. The distance between adjacent FP positions was 10 deg. The participant was instructed to follow the FP and maintain fixation for 1 s. After the calibration procedure, the participant pressed the response box to initiate a trial. The trial was started with an FP displayed at the centre of the monitor. If the participant maintained fixation for 1 s, the FP

disappeared and a testing image was presented at the centre of the screen. During the self-paced, free-viewing presentation, the participant was instructed to "judge the perceived image quality as accurately and as quickly as possible", and to respond by pressing a button on the response box with the dominant hand followed by a verbal report of subjective rating of the perceived image quality ranging from 1-7 (1 representing poor image quality and 7 excellent image quality). No reinforcement was given during this procedure.

The software developed in Matlab computed the recorded horizontal and vertical eye displacement signals as a function of time to determine eye velocity and position. Fixation locations were then extracted from the raw eye-tracking data using velocity (less than 0.2 deg eye displacement at a velocity of less than 20 deg/s) and duration (greater than 50 ms) criteria (Guo et al., 2006, 2012).

Results

Subjective quality rating analysis

To examine whether image distortion reduced subjective rating of the perceived image quality, a repeated-measures analyses of variance (ANOVA) was conducted with image manipulation as the independent variable, and quality rating score as the dependent variable. Greenhouse–Geisser correction was applied where sphericity was violated. The analysis demonstrated that compared to the original high quality images, introducing noise has significantly reduced the perceived image quality ratings ($F(6, 138) = 23.78$, $p < 0.001$; Fig. 3).

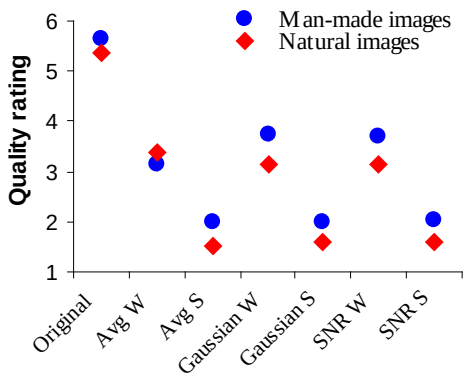

Figure 3. Subject quality rating for original man-mad and natural scene images and their distorted variants. Error bars represent standard error of mean.

A 3 (noise type: AVG, Gaussian, SNR) × 2 (noise intensity: weak, strong) × 2 (image type: man-made, natural) ANOVA was then conducted to examine to what extent different noise types and intensities affected the perceived image quality. The analysis revealed non-significant main effect of noise type ($F(1.31, 30.01) = 1.72$, $p = 0.19$), but significant main effect of noise intensity ($F(1, 23) = 715.06$, $p < 0.001$) and image type ($F(1, 23) = 55.87$, $p < 0.001$). It seems that strong noise affected image quality evaluation more than weak noise, but different noise types had the same deterioration impact on the perceived image quality. Interestingly, compared to man-made scenes, the same noise type and intensity (except for Avg W) led to slightly lower quality rating on natural scenes.

Gaze behaviour analysis

Fixation numbers: Compared to the original high quality images, participants directed fewer numbers of fixations when assessing the quality of distorted images ($F(6, 138) = 4.63$, $p < 0.001$; Fig. 4A). A 3 (noise type) × 2 (noise intensity) × 2 (image type) ANOVA with number of fixations per image as the dependent variable showed significant main effect of image type ($F(1, 23) = 15.83$, $p = 0.001$) with distorted man-made scenes attracting more fixations than natural scenes, and significant main effect of noise intensity ($F(1, 23) = 56.5$, $p < 0.001$) with higher intensity noise reducing the number of fixations directed at either man-made or natural scenes. There was also significant main effect of noise type ($F(2, 46) = 4.44$, $p = 0.02$), and interaction between image type and noise type ($F(2, 46) = 4.44$, $p = 0.004$). Specifically, images with SNR distortion tended to attract more fixations than those with Gaussian distortion, and except for Avg W other distortion types led to more fixations to man-made scenes than to natural scenes.

Viewing time: In general, original high quality images tended to attract longer viewing time than distorted images ($F(3.80, 78.67) = 5.84$, $p < 0.001$; Fig. 4B). A 3 (noise type) × 2 (noise intensity) × 2 (image type) ANOVA with viewing time per image as the dependent variable revealed non-significant main effect of image type ($F(1, 23) = 2.18$, $p = 0.15$), but significant main effect of noise intensity ($F(1, 23) = 31.19$, $p < 0.001$) with higher intensity noise shortening viewing time needed for image quality assessment, and significant main effect of noise type ($F(2, 46) = 12.49$, $p < 0.001$) with SNR distortion leading to longer image viewing time than Avg or

Gaussian distortion (all *ps* < 0.01). No difference in viewing time was observed between Avg and Gaussian distortion (p = 0.69).

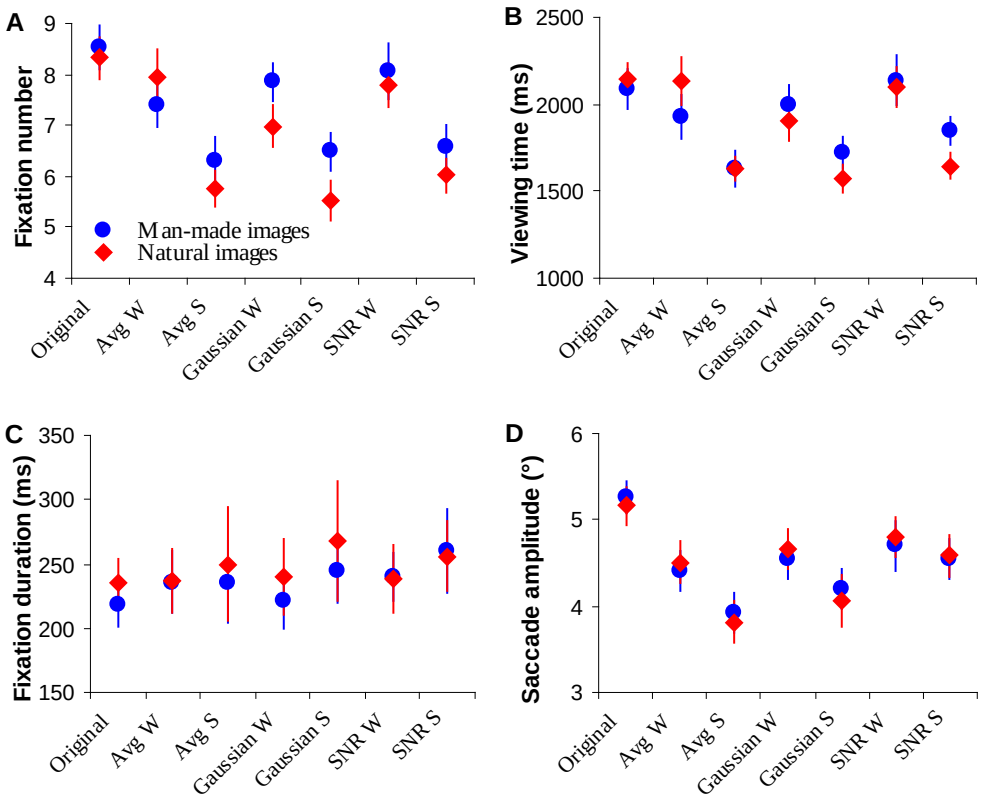

Figure 4. Number of fixations (A), viewing time (B), fixation duration (C) and saccade amplitude (D) associated with the evaluation of original man-mad and natural scene images and their distorted variants. Error bars represent standard error of mean.

Fixation duration: We then compared the average fixation duration across different image distortion conditions for all participants. As is normally the case with fixation duration data, the distributions were skewed with a majority of fixations lasting for a relatively short time, but a minority of long-lasting fixations. These outliers (those fixations differing more than two standard deviations from the median fixation duration) were discarded from this dataset. In total, less than 1% of fixation data was removed from further analysis. Compared to the distorted images, the average fixation duration was slightly shorter when assessing high quality images (F(6, 138) = 3.81, p = 0.002; Fig. 4C). A 3 (noise type) × 2 (noise intensity) × 2 (image type) ANOVA with fixation duration as the dependent variable demonstrated a significant main effect of noise intensity (F(1, 23) = 12.2, p = 0.002) with higher intensity noise leading to longer fixation duration, and significant main effect of image type (F(1, 23) = 9.3, p = 0.006) with natural scenes attracting slightly longer fixation duration than

man-made scenes. Noise type, on the other hand, had no marked impact on the fixation durations ($F(1.4, 31.7) = 0.31$, $p = 0.66$). The significant interaction between noise type and image type ($F(1.5, 34.5) = 4.3$, $p = 0.03$) further revealed that Gaussian distortion tended to induce shorter fixation duration for man-made scenes than for natural scenes.

Saccade amplitude: Image distortion also affected saccade amplitude in scene viewing with longer saccadic amplitude associated with high-quality images ($F(2.36, 45.52) = 13.92$, $p < 0.001$; Fig. 4D). A 3 (noise type) × 2 (noise intensity) × 2 (image type) ANOVA with saccade amplitude as the dependent variable demonstrated a significant main effect of noise intensity ($F(1, 23) = 7.13$, $p = 0.02$) with higher intensity noise leading to shorter saccade amplitude, and significant main effect of noise type ($F(2, 46) = 11.14$, $p < 0.01$) with SNR distortion inducing longer saccade amplitude than Avg or Gaussian distortion. Image type had no impact on saccade amplitude ($F(1, 23) = 0.15$, $p = 0.7$).

Fixation distribution: Finally to examine to what extent image distortion affected participants' fixation distribution over the images, we measured two metrics, fixation distance from the image center and entropy, to quantify the difference between the spread of the fixations across different fixation maps (Judd et al., 2011). As shown in Fig. 5A, participants demonstrated stronger central bias (i.e. fixating at local regions close to image centre) when viewing degraded scenes. A 3 (noise type) × 2 (noise intensity) × 2 (image type) ANOVA with fixation distance from the image centre as the dependent variable demonstrated a significant main effect of noise type ($F(1.5, 34.5) = 17.9$, $p < 0.001$), noise intensity ($F(1, 23) = 137.8$, $p < 0.001$) and image type ($F(1, 23) = 7.0$, $p = 0.02$). Clearly, compared to man-made scenes, participants inspected more at the central regions of natural scenes. Regardless of image type, the average distance of fixation from the image centre decreased with increased noise intensity, and Avg and Gaussian distortion induced more centred fixations than SNR distortion.

The analysis of entropy data further showed a more widely spread fixation distribution (reflected by higher entropy value) over original high quality images than the degraded images (Fig. 5B). A 3 (noise type) × 2 (noise intensity) × 2 (image type) ANOVA with entropy value as the dependent variable revealed significant main

effect of noise intensity ($F(1, 23) = 137.8$, $p < 0.001$) with higher intensity noise resulting in spatially more restricted fixation distribution over the image, and significant main effect of noise type ($F(1.6, 36.8) = 30.8$, $p < 0.001$) with SNR distortion inducing higher entropy values than Avg or Gaussian distortion. Image type, on the other hand, had no significant impact on entropy value ($F(1, 23) = 2.3$, $p = 0.14$).

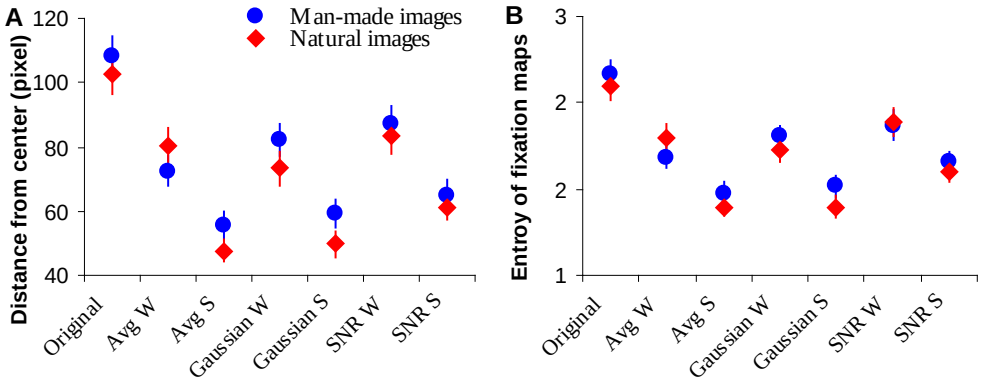

Figure 5. Comparison of fixation distribution in viewing of original man-made and natural scene images and their distorted variants, using metrics of fixation distance from the image centre (A) and entropy of the fixation map (B). Error bars represent standard error of mean.

Experiment 2: Computational modelling

Our eye-tracking study in Experiment 1 has clearly demonstrated that adding noise into scenes would significantly decrease the perceived image quality and affect gaze behaviour associated with the task of image quality assessment. In comparison with original high quality images, distorted images with decreasing quality gradually attracted fewer numbers of fixations but longer fixation durations, shorter saccades and less scatter of fixation positions around the image centre. Interestingly, this systematically varied gaze behaviour in scene viewing was mainly determined by the perceived image quality (or noise intensity) although noise type and image category also played a role.

In Experiment 2, we aimed to examine how well the state-of-the-art computational models of visual attention could predict human gaze allocation in viewing of degraded scenes. We selected only the most sophisticated and best-performing models in previous benchmarking tests, including Judd model (Judd et al., 2009), Erdem model (Erdem & Erdem, 2013), Graph-based visual saliency (GBVS)

model (Harel, Koch, & Perona, 2007), and Adaptive whitening saliency (AWS) model (Garcia-Diaz et al., 2012a, 2012b). Fig. 6 shows examples of four high quality scene images used in Experiment 1 and the corresponding saliency maps generated by the chosen models.

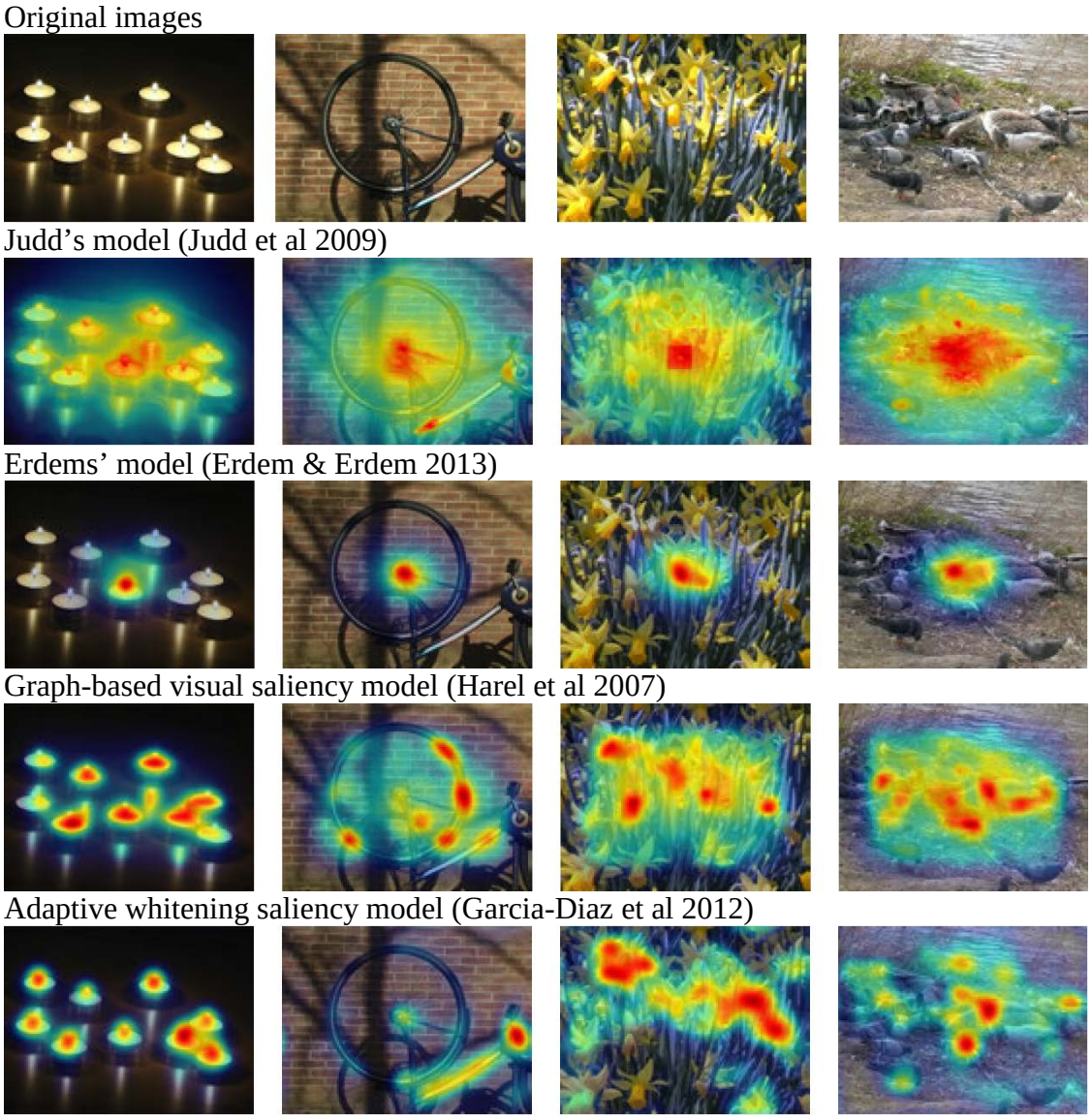

Figure 6. Saliency maps generated by different computational models of visual attention, and projected onto the original images in the form of heat maps.

Judd model (Judd et al., 2009) has introduced a machine learning approach to achieve an optimal combination of feature maps computed by the bottom-up saliency model (Itti & Koch, 2000). It also takes global scene context into account, and includes object/person detection and central bias feature. This model has the highest

score in a benchmark test by Judd et al. (2012). Erdem model is a recently published model emphasizing on the optimal, non-linear combination of features using region covariance matrices (Erdem & Erdem, 2013). It also incorporates a central bias similar to the center feature of Judd model. Using the same benchmarking image set (Judd, Durand, & Torralba, 2012), this model has achieved a similarly high score as Judd model. GBVS model (Hare, Koch, & Perona, 2007) works on similar image features to those by Itti and Koch (2000), and detects neighbouring feature differences. It then forms and normalizes activation maps on certain feature channels to highlight conspicuity. In contrast to Judd and Erdem model, GBVS model has an intrinsic centre bias which is not part of any feature, but intrinsic to the construction of the saliency map. This model scored very well in two benchmark tests by Judd et al. (2012) and Borji et al. (2013). AWS model is essentially based on the idea of adapting the basis of the low-level features to the specific statistical structure of the image (Garcia-Diaz et al., 2012a, 2012b). This model does not contain any form of central bias. Nevertheless, it is capable of predicting gaze locations with high accuracy. In contrast to the results by Judd et al. (2012), Borji et al. (2013) found the AWS model to be the best performing model using a different scoring method.

Interestingly, in many cases a simple bottom-up saliency map with a Gaussian blur around the image centre (centre bias model) can yield comparatively good results and perform on the same level as the high scoring models described above (Judd, Durand, & Torralba, 2012; Borji, Sihite, & Itti, 2013). This fits well with human natural gaze behaviour of central fixation bias in scene viewing ─ an effect that can be ascribed to several factors. First, with photography, objects of interest are most often placed in the centre of the image (Zhang et al., 2008; Tseng et al., 2009). Second, the image centre as point of fixation could ensure rapid access to every point of interest in the image. It might therefore be of advantage to start scene viewing in the centre (Tatler, 2007). The motor bias, tendency to make more shorter and horizontal saccades, does not play a big role in central fixation bias. It is therefore not problematic to let subjects fixate the centre location before image presentation (Tseng et al., 2009). While this centre bias model is certainly valuable in the context of computer and television screens, its validity on unrestricted viewing behaviour in natural conditions (e.g., in an open environment) might be questionable.

Methods

Performance measurement

For each tested image, its saliency map (or predicted fixation map) was computed by Judd model, Erdem model, GBVS model, and AWS model separately using Matlab programmes obtained from the model developers. To systematically compare the predictive power of these models, we used four different measures of performance that are common in literature.

The Receiver Operating Characteristic (ROC) score or the area under the ROC curve (AUC) is arguably the most common metric to evaluate how well a computed saliency map predicts actual human fixation map. The saliency map is treated as a binary classifier on every pixel (either fixated or not fixated) in the image. The ROC curve can be drawn by varying the classifier's threshold in percent of the image pixels being classified as fixated. The AUC allows determining the performance of the classifier for different thresholds. The ROC score can be interpreted as the probability that an actual fixation location is ranked more highly than a non-fixated location for the given saliency map. Chance performance yields a ROC score of 0.5 and the optimal performance is 1 (Judd, Durand, & Torralba, 2011).

As ROC score is based on the rank of the fixations and not on absolute metric differences, a high number of true positives will lead to a high ROC score regardless of false alarm rate. It is therefore argued that the ROC score alone is not sufficient to fully evaluate a model's predictive power (Zhao & Koch, 2011). Given this consideration, two additional performance measurements were implemented. The Earth Mover's Distance (EMD, Rubner, Tomasi, & Guibas, 2000) takes into account the absolute distance rather than rank, and measures the cost that is necessary to transform one fixation map into another. The larger the EMD the less similar are the two fixation map distributions, while an EMD close to zero indicates that the two distributions are very much alike. Similarity score (Judd, Durand, & Torralba, 2012) also compares two fixation maps with one indicates two identical distributions, and zero indicates two completely different distributions. Importantly, the saliency maps have to be comparable in terms of general brightness for calculating EMD and similarity score. As saliency maps from different models vary greatly in the amount of returned salient pixels, the histograms of the saliency models have been matched to the human fixation map before computing both EMD and similarity scores (Judd et al., 2009).

The similarity score can also compare two saliency maps directly. This measurement is particularly useful for the current experimental design as we obtained saliency maps for the same image with different types and amount of noise added. We investigated the change in saliency maps that resulted from the application of noise using the similarity score, testing every noise condition against the original image as baseline.

As the centre bias often plays an important role for the success of visual attention models, we also measured shuffled ROC score that is more sensitive to off-centre fixations which are usually harder to predict (Borji, Sihite, & Itti, 2013). Zhang et al. (2008) computed this score by choosing a different set of negative fixation locations (see also Tatler et al., 2005). While the ROC score chooses all non-fixated image pixels as a negative sample set, shuffled ROC score uses negative sample set consisting of all fixations of a subject on the other images. In our case, the negative sample includes 9 fixation maps from the same participant on other images belonging to the same image type and distorted with the same noise type and intensity.

Human baseline performance

The baseline of human performance is determined using the ROC score, which measures how well the fixations of each participant can be predicted by those of other participants (all-except-one observers). Specifically, we select one participant's fixation map as actual fixations that we wanted to predict. Instead of creating a normal saliency map based on image features, we used the fixations of the remaining participants to create a simple saliency map by marking all the fixated positions in the image and convolving a Gaussian over these locations (Judd, Durand, & Torralba, 2011). The size of the Gaussian has a cut-off frequency of 8 cycles per image, corresponding to about one degree of visual angle. By repeating this procedure for all participants and averaging the resulting ROC scores, we obtained a measure for the variability (or inter-observer agreement) within human gaze patterns that can serve as an upper bound to the performance of a given computational model.

We used a simple script provided by Judd et al. (2012) to compute the saliency map for the centre bias model. It creates a symmetric Gaussian blob at the image centre, which is stretched horizontally in order to fill the image completely. For many datasets, the stretched version of the Gaussian blob performs slightly better than an

isotropic Gaussian, and can reach up to 0.8 ROC score (Zhang et al., 2008; Judd, Durand, & Torralba, 2012).

Results

Analysis of model performance

Given Experiment 1 has shown that human gaze behaviour in scene viewing was mainly determined by the noise intensity, in Experiment 2 we focused on the general impact of noise intensity on the model performance and collapsed different noise types into one group to simplify data analysis. 4 (model type: Judd model, Erdem model, GBVS model, AWS model) × 3 (noise intensity: no-noise, weak, strong) × 2 (image type: man-made, natural) ANOVAs were then performed with the computed ROC score, similarity score, EMD score, and shuffled ROC score as the dependent variables.

ROC score: to compare the models' performance relative to the human baseline, we computed a normalized AUC score as quotient of model performance and human baseline performance (Fig. 7). The analysis revealed non-significant main effect of noise intensity ($F(2, 542) = 0.04$, $p = 0.96$), but significant main effect of model type ($F(3, 542) = 195.7$, $p < 0.001$) and image type ($F(1, 542) = 12.1$, $p < 0.001$). There was also significant interaction between model type and image type ($F(3, 542) = 21.5$, $p < 0.001$) and between model type and noise intensity ($F(6, 542) = 3.6$, $p = 0.002$). The interaction between image type and noise intensity was not significant ($F(2, 542) = 0.05$, $p = 0.95$).

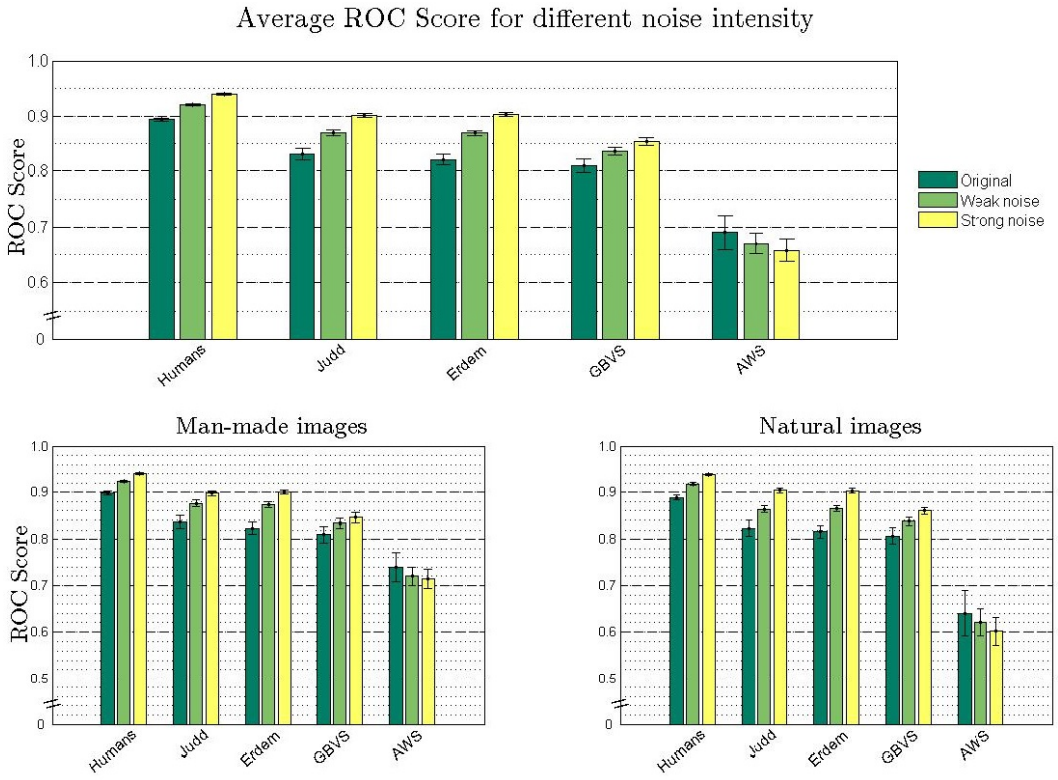

Figure 7. Average ROC scores from different visual attention models in computing saliency map of man-made and natural scene images with different noise intensities. The upper panel shows the aggregated data over image types, and the lower panel shows ROC scores separately for the two image types. Error bars represent standard deviation.

Across individual models, there was no significant difference in performance between Judd and Erdem models ($p > 0.05$), but GBVS model performed worse than Judd and Erdem models, and AWS model had the lowest score compared with other models (all $ps < 0.05$). Furthermore, for Judd, Erdem and GBVS models, increasing noise intensity led to the increased ROC scores (all $ps < 0.05$).

Although Judd and Erdem models performed very well when measured with ROC score, they still did not reach the levels of human performance. The trivial centre bias model performed almost equally well. This result is in line with previous research showing the marked impact of the centre bias on all saliency models predicting human fixations (Judd, Durand, & Torralba, 2012; Borji, Sihite, & Itti, 2013). Only AWS model does not incorporate the centre bias and consequently performed significantly worse. AWS model also performed much worse on natural scenes compared to man-made scenes. For the other three models this effect was less pronounced.

Furthermore, the higher ROC scores of human baseline performance in Fig. 7 indicated limited individual differences in gaze behaviour. A 2 (image type: man-made, natural scene) × 3 (noise intensity: no-noise, weak, strong) ANOVA revealed a significant main effect of noise intensity ($F(1.3, 30.0) = 70$, $p < 0.001$), but non-significant effect of image type ($F(1, 23) = 2.8$, $p = 0.11$) and non-significant interaction between image type and noise intensity ($F(2, 46) = 0.96$, $p = 0.39$). Specifically, the human baseline was not affected by image type. Noise intensity however did have a strong impact; increasing noise intensity would reduce the variation of fixation distribution between individual participants.

Similarity score: The 4×3×2 ANOVA showed significant main effect of noise intensity ($F(2,542) = 136$, $p < 0.001$; Fig. 8) and model type ($F(3, 542) = 119.5$, $p < 0.001$), but non-significant main effect of image type ($F(1,542) = 2.9$, $p = 0.88$) and all the interaction effects between independent variables (all *ps* > 0.15). For this performance measurement, all models scored worse when increasing noise intensity to the images. Furthermore, Judd model performed significantly better than the other three models over all conditions (all *ps* < 0.05). The other models did not differ significantly from each other (all *ps* > 0.05).

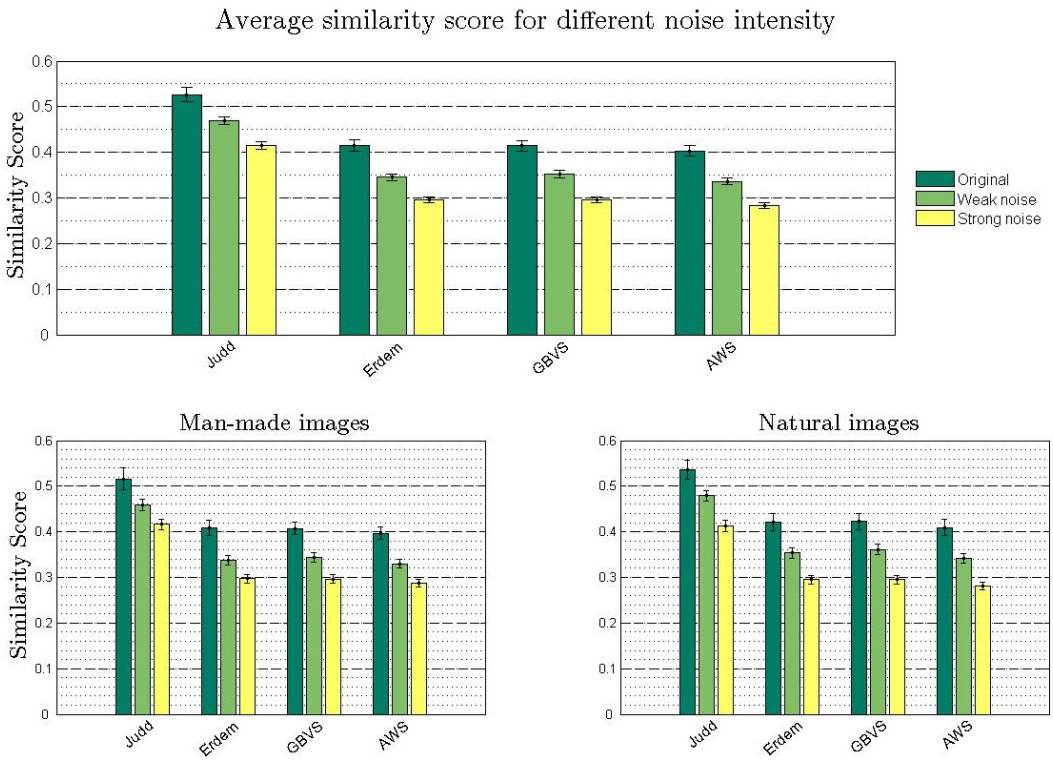

Figure 8. Average similarity scores from different visual attention models in computing saliency map of man-made and natural scene images with different noise intensities. The upper panel shows the aggregated data over image types, and the lower panel shows similarity scores separately for the two image types. Error bars represent standard deviation.

EMD score: Contrary to ROC and Similarity scores, a low EMD score signals high performance for a model. The 4×3×2 ANOVA revealed significant main effect of model type ($F(3, 542) = 259.6$, $p < 0.001$; Fig. 9), noise intensity ($F(2, 542) = 3.4$, $p = 0.03$) and image type ($F(1, 542) = 5.8$, $p = 0.02$). Specifically Judd and Erdem models performed significantly better than GBVS and AWS models, and AWS model showed the poorest performance (all *ps* < 0.05). The significant interaction between model type and image type ($F(3, 542) = 11.3$, $p < 0.001$) further revealed that AWS model performed sharply worse for natural scenes, while Judd and Erdem models performed equally well for man-made and natural scenes. There was also a significant interaction between model type and noise intensity ($F(6, 542) = 9.3$, $p < 0.001$) with AWS model performing significantly worse with increased noise intensity. Judd and Erdem models, on the other hand, showed better performance with increased noise intensity (all *ps* < 0.05). The interaction between image type and noise intensity was not significant ($F(2, 542) = 0.09$, $p = 0.91$).

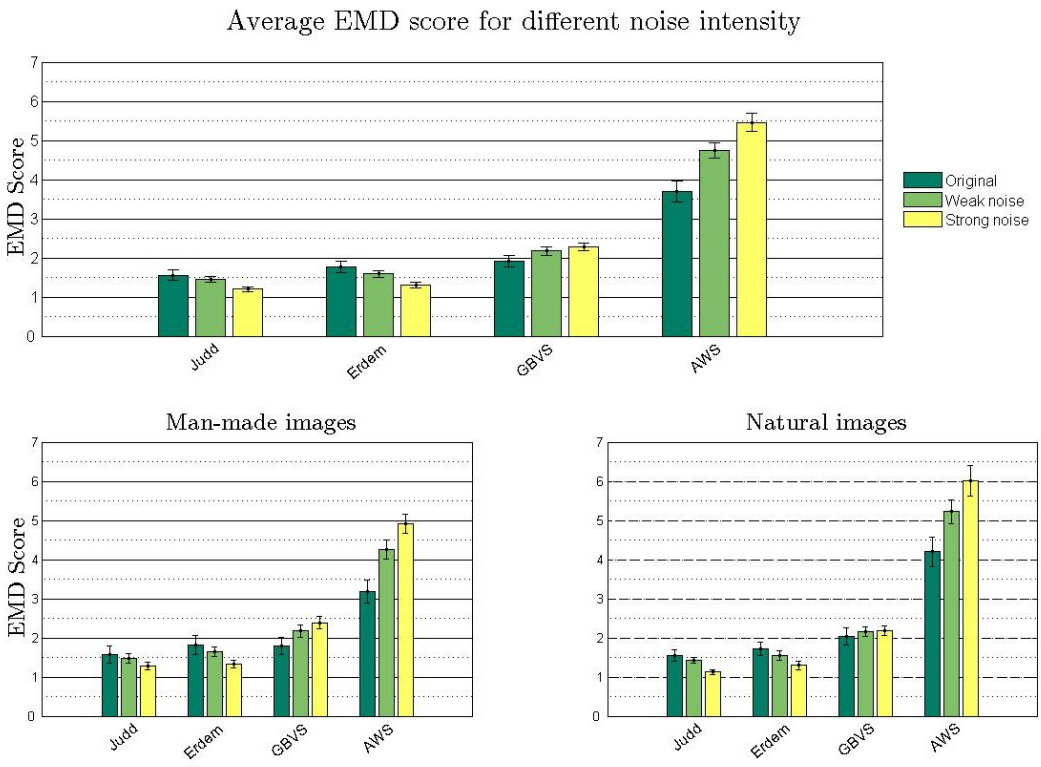

Figure 9. Average EMD scores from different visual attention models in computing saliency map of man-made and natural scene images with different noise intensities. The upper panel shows the aggregated data over image types, and the lower panel shows EMD scores separately for the two image types. Error bars represent standard deviation.

Shuffled ROC score: this score was computed to minimize possible artefacts from the centre bias and image borders. As shown in Fig. 10, the result patterns from other measurements were reversed when using shuffled ROC scores. The 4×3×2 ANOVA demonstrated significant main effect of model type ($F(1.1, 26.4) = 36.6$, $p < 0.001$) with AWS model showing the best performance followed by GVBS and Judd models, and finally by Erdem model. There was also significant main effect of noise intensity ($F(1.3, 28.8) = 5.4$, $p = 0.02$) with decreased model performance associated with increased noise intensity. The non-significant main effect of image type ($F(1, 23) = 0.76$, $p = 0.79$) but significant interaction between model type and image type ($F(2.1, 48.8) = 10.5$, $p < 0.001$) indicated that AWS model performed markedly worse for natural scenes than for man-made scenes.

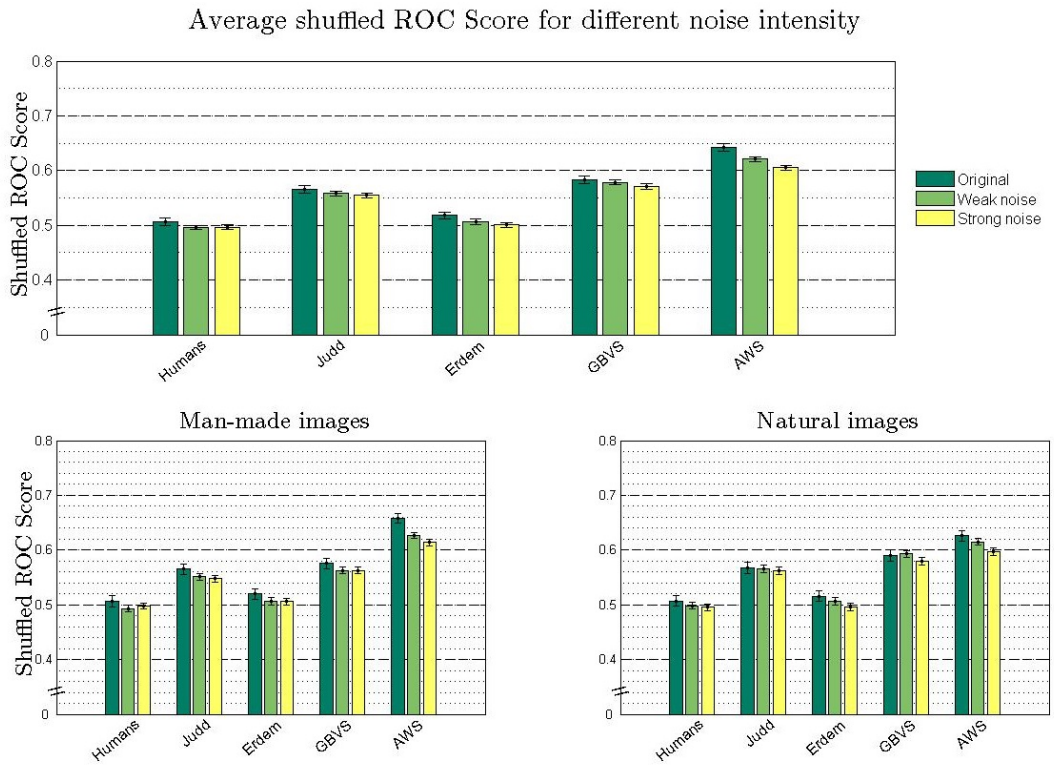

Figure 10. Average shuffled ROC scores from different visual attention models in computing saliency map of man-made and natural scene images with different noise intensities. The upper panel shows the aggregated data over image types, and the lower panel shows shuffled ROC scores separately for the two image types. Error bars represent standard deviation.

Effect of noise type on saliency maps

To directly examine the impact of different noise types and intensities on saliency maps computed by a given model, for each model and each image we computed the similarity score between saliency maps from original high-quality image and from noisy versions of the same image (Fig. 11). Judd (M = 0.98) and GBVS models (M = 0.99) performed almost equally well under different noise types and intensities, followed by Erdem model (M = 0.97). AWS model (M = 0.92) performed significantly lower than the other three models (SEM was 0.003 for all models).

Interestingly, for actual human fixation maps, increasing noise intensity has clearly decreased similarity between actual fixation distributions in high-quality and degraded images (Post-hoc test for each noise type, all *ps* < 0.05). However, at a given noise intensity, such change in similarity was not affected by the noise type (all *ps* > 0.05); suggesting that human fixation map in scene viewing was more sensitive

to noise intensity rather than noise type. This effect was not clear for predicted fixation maps by the tested models. In comparison with human performance, these visual attention models were either not sensitive to noise type and noise intensity (e.g., Judd and GBVS models) or overly sensitive to noise manipulations (e.g., AWS model, Fig. 11). In comparison with other noise types and intensities, high intensity SNR noise (SNR S, white Gaussian noise with a signal-to-noise ratio of 0) significantly decreased similarity scores computed by GBVS and AWS models (all *ps* < 0.05).

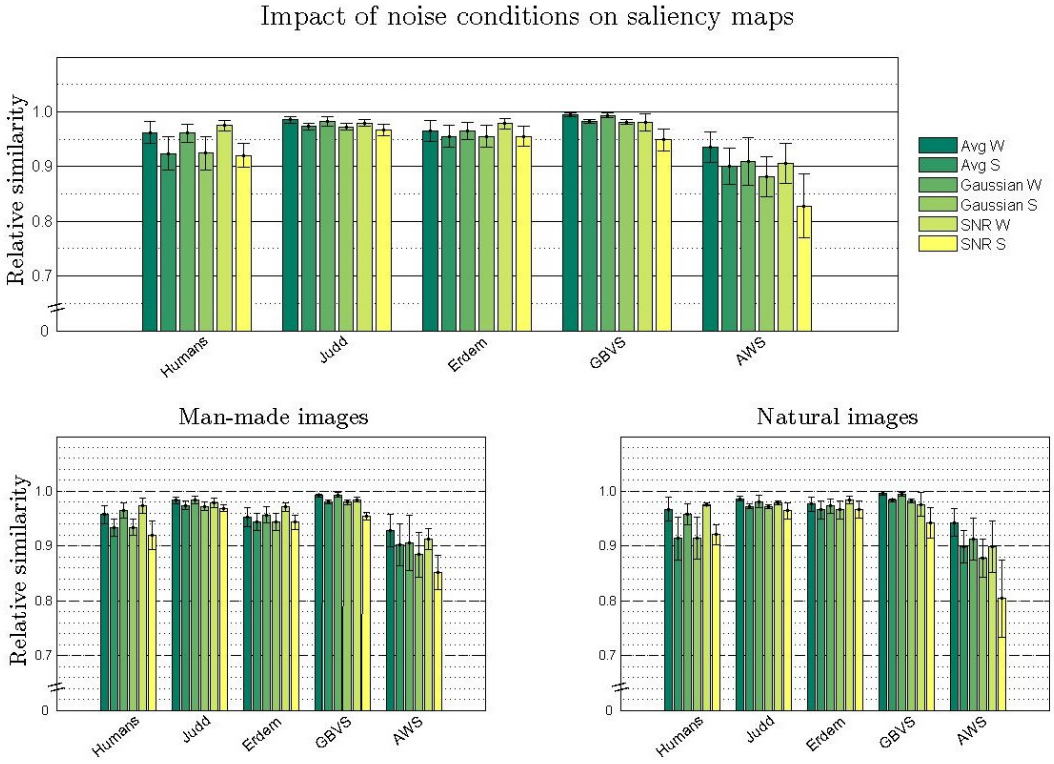

Figure 11. Changes in saliency map over different noise manipulations. For each model and each image we computed the similarity score between saliency maps of original high-quality image and noisy versions of the same image. Baseline refers to human performance. Error bars represent standard deviation.

Furthermore, we compared the impact of different noise manipulations on human fixation distributions using the computed similarity scores. The prediction matrix in Fig. 12 showed how human fixation distribution in one condition (described by the column entry) can be predicted by fixation distribution in another condition (rows). For instance, each entry in the first row demonstrated how well human fixation distributions in all noise conditions can be predicted using only the fixation

distribution in original high-quality images. The entries on the diagonal cells showed how strong the deviation of fixation distribution was, i.e. how well a saliency map based entirely on the fixations of all human observers predicts the fixations of a single human observer. As a baseline, we also included the performance of how well the centre model predicts fixations on each noise condition. From this prediction matrix, it is evident that fixations on the degraded images were poor predictors for fixations on the original high-quality images (i.e. first column in the matrix), but fixations on a degraded image could be very well predicted by fixations on images of different noise manipulations. It is also evident that fixations for an image at specific noise intensity were best predicted by fixations on the image at the same noise intensity regardless of noise types. In addition, actual human fixations provided better predictions than the centre model.

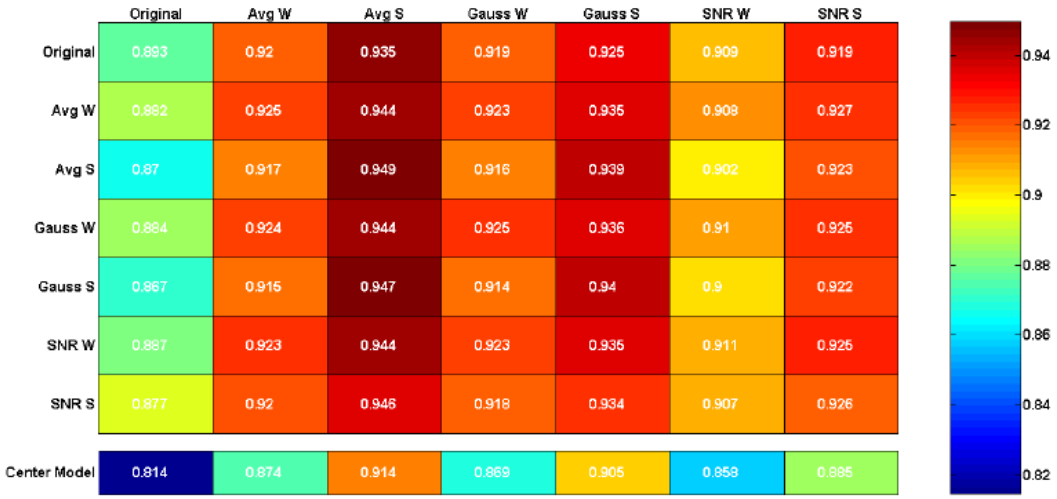

Figure 12. Matrix showing how well fixation distribution in one noise condition predicts the fixation distribution in another condition. For every condition a saliency map is created from human fixation distributions and is used to predict fixation distribution for either the same or a different noise condition. The saliency map is based on the averaged fixations in the conditions labelling the rows of the matrix. The fixations used to evaluate the saliency map classifier are taken from the conditions labelling the columns of the matrix.

Discussion

Human gaze behaviour in evaluation of noisy images

In this study we have demonstrated that introducing external image noise would significantly decrease the perceived image quality and affect gaze behaviour associated with the task of assessing image quality. In comparison with high quality

images, noisy images attracted fewer numbers of fixations but longer fixation durations, shorter saccades and stronger central fixation bias (Fig. 4 and 5). Probably due to different image statistics between natural and man-made scenes (e.g., more steady and uniform structure in natural scenes), the impact of noise manipulation is more pronounced for natural scenes than for man-made scenes.

It should be noted that the link between image noise and changes in saccade behaviour has been reported before. For instance, by applying masking, low- or high-pass spatial frequency filters and reducing image resolution to image region in fovea (van Diepen & d'Ydewalle, 2003), in periphery (Pomplun, Reingold, & Shen, 2001; Loschly & McConkie, 2002; Nuthmann, 2013) or in entire visual field (Mannan, Ruddock, & Wooding, 1995; Judd, Durand, & Torralba, 2011), previous studies have observed that these image manipulations could reduce saccadic selectivity and lead to shorter saccades and longer fixation durations or viewing time (see also van Diepen & Wampers, 1998). It is possible that noisy images could reduce the saliency of peripheral visual features and increases the difficulty of saccade target selection (Reingold & Loschky, 2002). The computational model of fixation duration proposed by Nuthmann et al. (2010) has indicated that task difficulty in visual and cognitive processing will inhibit saccade initiation and lead to longer fixation durations.

It has been well documented that early image viewing is often associated with a central fixation bias (Parkhurst et al., 2002; Tatler et al., 2005, 2007; Tseng et al. 2009). We observed this central bias was stronger when increasing image noise intensity (Fig. 5). Judd et al. (2011) also noticed an evident central fixation bias in low-resolution man-made scenes, and proposed that observers do not have to move their eyes away from the centre because they can resolve the entire low resolution image using peripheral vision and therefore gaze locations are more restricted to the centre. Our findings, however, may provide an alternative interpretation. Specifically, one could argue that both decrease in image resolution and increase in image noise will reduce the overall saliency from out-of-focus regions and make saccade initiation to peripheral regions more difficult, resulting in fewer fixations to non-centre areas. Indeed, the saliency maps computed by visual attention models demonstrated a decreased overall image saliency with increasing noise intensities (Fig. 11).

The increased central fixation bias in degraded scenes has led to an increase in inter-observer consistency in gaze allocation, which has been reflected by a lower inter-observer variation in fixation maps for noisy images compared to those for

original high quality image. As demonstrated in Fig. 12, noise intensity was the major factor, whereas different noise types had only a minor effect on the degree of central bias and inter-observer gaze consistency. These observations would have important implication on the evaluation of model performance as different models incorporate and weigh central bias differently (Borji, Sihite, & Itti, 2013).

Given the single task (judging image quality) design adopted in the current study, it is difficult to be certain whether the observed changes in gaze pattern in viewing degraded scenes was due to changes in cognitive strategy to complete the task or changes in added noise type and intensity. Considering that similar relations between image noise and changes in saccade behaviour has been reported in previous studies using different task-demands (e.g., free-viewing, image identification, object search) (Mannan, Ruddock, & Wooding, 1995; Judd, Durand, & Torralba, 2011; Nuthmann, 2013), and our data collection had consistent task-demand and randomised brief image presentation (drawn from different image categories, noise types and intensities); we speculated that the observed gaze pattern changes in scene viewing were more likely due to the noise intensity rather than the task strategy. However, as cognitive demand can affect our gaze behaviour in natural vision (Tatler et al., 2011), it remains to be seen to what extent the current findings can be generalised to different cognitive processes, such as scene understanding and recognition.

Model performance in evaluation of noisy images

Two recent studies have systematically compared the performance of visual attention models in predicting human gaze allocation. Judd et al. (2012) found that Judd and Erdem models (see http://people.csail.mit.edu/tjudd/SaliencyBenchmark/index.html for up-to-date results including Erdem model) had the best predicting performance. In contrast, Borji et al. (2013) argued that metrics used by Judd et al. (2012) for performance evaluation – ROC, similarity and EMD score – were to some extent influenced by central fixation bias, and proposed shuffled ROC score which was neutral to centre bias. They consequently found that AWS model had the best predictive power.

In our study, Judd, Erdem and GBVS models performed equally well if we aggregated their performance over all noise conditions on ROC score. The good performance of Erdem model was particularly remarkable given the limited number of features used by this model. However, no model has achieved a performance

anywhere close to human level, so there is still room for improvement. AWS model performed considerably worse than other three models, probably because it doesn't use any location related information. On the other hand, the high predictive power of both Judd and Erdem models could be largely attributed to the centre bias factor incorporated into them. In fact, the performance of these two models without centre bias was only slightly better than that of AWS model. As GBVS model incorporates centre bias implicitly, we could not compute any saliency map without a centre bias for this model. But given the good performance of GBVS, Judd and Erdem models, it seems that incorporating centre bias renders an advantage to visual saliency models, at least when using the standard ROC score.

All models performed distinctly worse for natural scenes than for man-made scenes, even though this effect was somewhat mitigated in models incorporating centre bias. This might be due to different image statistics between natural and man-made images, in particular the steady and uniform structure in natural scenes. For all models except AWS model, increasing noise intensity led to a slight increase in ROC score compared with no noise condition. This effect could be partly attributed to the increased role of centre bias, which is in line with the actual human gaze behaviour change (increased central fixation bias with increased noise intensity). The reversed pattern of AWS model (increasing noise intensity leading to worse model performance) was possibly due to the fact that this model is based on statistical properties of the image. The noise manipulation in our experiment may have changed image statistics in a particularly adverse manner for AWS model.

We found that similarity score was less informative to compare performance between different models. Although showing a slight lead for Judd model, it was unable to differentiate between Erdem, GBVS and AWS models (Fig. 7, 8, 9 and 10), suggesting that using this metric alone to evaluate visual attention model performance is insufficient and the practical significance of this metric is limited.

The evaluation with EMD score, on the other hand, reached similar conclusion as those with ROC score. Specifically, Judd and Erdem models had the best performance, followed by GBVS model, and then by AWS model. The enhanced performance in Judd and Erdem model was largely due to centre bias incorporated in these models. Nevertheless, the Judd and Erdem model without centre bias still performed slightly better than the AWS model. Importantly, similar to ROC score, the

noise intensity has opposite impact on Judd and Erdem models in comparison with AWS and GBVS models.

In order to adequately account for the important role of centre bias in our evaluation, we also computed shuffled ROC score as proposed by Borji et al. (2013). This metric was not only ignoring centre bias but effectively penalizing those models incorporating some form of centre bias (Fig. 7, 8, 9 and 10). As a consequence, human baseline and centre model performance were reduced to chance level (~0.5 shuffled ROC score) and those models incorporating centre bias performed significant worse using this metric. Considering this penalty, AWS model performed slightly better than GBVS, Judd and Erdem models. Based on these findings, especially poor performance for human baseline, we propose that assessing model performance purely based on shuffled ROC score is inadequate. Furthermore, our results clearly demonstrated that GBVS model, which performed very well in the study by Borji et al. (2013) using NSS and CC scores, performed distinctly worse for shuffled ROC score in this experiment. A similar adverse effect may have substantially influenced performance of many other models incorporating the centre bias factor.

Biological plausibility of visual attention models

Our experimental design allowed us to use similarity score to directly examine how human gaze behaviour or predictive performance from a given visual attention model was affected by different noise types and noise intensities. As shown in Fig. 11, human fixation map was significantly modulated by noise intensity but not by noise type. The results from visual attention models, however, were different from human responses. While Judd and Erdem models were rather insensitive to both noise types and noise intensities, the AWS model was overly sensitive to noise manipulation, suggesting these visual attention models lacked human-like sensitivity to noise type and intensity, and could not represent the processing of image noise in human visual system.

For instance, compared with other noise types and intensities, high intensity SNR noise exerted a strong influence on the saliency map computed by AWS model. On the other hand, human gaze behaviour was not affected differently by SNR noise in comparison with other noise types. It seems that there must be either some underlying mechanism in humans that can correct for distortion by SNR noise, or alternatively the visual system processes incoming visual input in a way that is more

robust against this particular noise. These results indicate that AWS model does not reflect the actual way of information processing in human visual system, even though it can explain some other psychophysical and perceptual observations (Garcia-Diaz et al., 2012a, 2012b).

Conclusion

In this study we observed that image noise consistently affected human gaze behaviour. The impact was strongly dependent on noise intensity, while effects between different noise types were only minor. Importantly, our results showed the increasing importance of central fixation bias for interpreting fixation distribution on the degraded scenes, which has direct implications for the construction and implementation of visual attention models in technical applications. We also found that estimating model performance depended critically on the choice of evaluation metric, and in particular on whether it factored in a centre bias. Further improvements in model predictive power might be fostered by biological insights into the exact functioning of human visual system, such as robustness and adaptability to external noises.

References

Acik, A., Onat, S., Schumann, F., Einhäuser, W., & König, P. (2009). Effects of luminance contrast and its modifications on fixation behaviour during free viewing of images from different categories. *Vision Research, 49*, 1541–1553.

Allard, R., & Cavanagh, P. (2012). Different processing strategies underlie voluntary averaging in low and high noise. *Journal of Vision, 12(11):6*, 1–12.

Betz, T., Kietzmann, T. C., Wilming, N., & König, P. (2010). Investigating task-dependent top-down effects on overt visual attention. *Journal of Vision, 19(3):15,* 1-14.

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modelling: a comparative study. *IEEE Transactions on Image Processing, 22*, 55-69.

Castelhano, M. S., & Henderson, J. M. (2008). The influence of color on perception of scene gist. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 660–675.

Einhäuser, W., Rutishauser, U., Frady, E. P., Nadler, S., König, P., & Koch, C. (2006). The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision, 6,* 1148–1158.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8(14):18,* 1–26.

Erdem, E., & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision, 13(4):11,* 1–20.

Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012a). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing, 30,* 51–64.

Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012b). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision, 12(6):17,* 1–22.

Guo, K., Mahmoodi, S., Robertson, R. G., & Young, M. P. (2006). Longer fixation duration while viewing face images. *Experimental Brain Research, 171,* 91–98.

Guo, K., Smith, C., Powell, K., & Nicholls, K. (2012). Consistent left gaze bias in processing different facial cues. *Psychological Research, 76,* 263-269.

Goferman, S., Zelnik-Manor, L., & Tal, A. (2012). Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34,* 1915-1926.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advance in Neural Information Processing Systems (NIPS),* pp. 545–552. Cambridge, MA: MIT Press.

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science, 16,* 219-222.

Isaacowitz, D. M. (2006). Motivated gaze: The view from the gazer. *Current Directions in Psychological Science, 15,* 68–72.

Itti, L. & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *IEEE 12th International Conference on Computer Vision,* 2106-2113.

Judd, T., Durand, F., & Torralba, A. (2011). Fixations on low-resolution images. *Journal of Vision 11(4):14,* 1–20.

Judd, T., Durand, F., & Torralba, A. (2012). *A benchmark of computational models of saliency to predict human fixations (Tech. Rep. No. MIT-CSAIL-TR-2012-001)*. Cambridge, MA: MIT Computer Science and Artificial Intelligence Laboratory.

Kanan, C., Tong, M. H., Zhang, L., & Cottrell, G. W. (2009). SUN: Top-down saliency using natural statistics. *Visual Cognition, 17,* 979–1003.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision, 13,* 201-214.

Loschky, L. C., & McConkie, G. W. (2002). Investigating spatial vision and dynamic attentional selection using gaze-contingent multi-resolutional display. *Journal of Experimental Psychology: Applied, 8,* 99-117.

Mannan, S., Ruddock, K., & Wooding, D. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-d images. *Spatial Vision, 9,* 363–386.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision, 10,* 165–188.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45,* 205–231.

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). Crisp: A computational model of fixation durations in scene viewing. *Psychological Review, 117,* 382-405.

Nuthmann, A. (2013). On the visual span during object search in real-world scenes. *Visual Cognition, 21,* 803-837.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107–123.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16,* 125–154.

Pollux, P. M. J., Hall, S., & Guo, K. (2014). Facial expression training optimises viewing strategy in children and adults. *PLoS ONE, 9(8),* e105418.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems, 10,* 341–350.

Reingold, E. M., & Loschky, L. C. (2002). Saliency of peripheral targets in gaze-contingent multiresolutional displays. *Behavior Research Methods, Instruments, & Computers, 34,* 491-499.

Péteri, R., Fazekas, S., & Huiskes, M. J. (2010). DynTex: A comprehensive database of dynamic textures. *Pattern Recognition Letters, 31,* 1627-1632.

Pomplun, M., Reingold, E. M., & Shen, J. (2001). Peripheral and parafoveal cueing and masking effects on saccadic selectivity in a gaze-contingent window paradigm. *Vision Research, 41,* 2757-2769.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision, 40,* 99–121.

Sheikh, H. R., Bovik, A. C., & de Veciana, G. (2005). An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing, 14,* 2117-2128.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45,* 643–659.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7(14):4,* 1–17.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11(5):5,* 1–23.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113,* 766–786.

Torralba, A. (2009). How many pixels make an image? *Visual Neuroscience,* 26, 123–131.

Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9;9(7):* 4.

van Diepen, P. M., & Wampers, M. (1998). Scene exploration with Fourier-filtered peripheral information. *Perception, 27,* 1141-1151.

van Diepen, P., & d'Ydewalle, G. (2003). Early peripheral and foveal processing in fixations during scene perception. *Visual Cognition, 10,* 79-100.

Watson, A. B., & Ahumada, A. J. (2011). Blur clarified: A review and synthesis of blur discrimination. *Journal of Vision, 11(5):10,* 1–23.

Winkler, S. (2012). Analysis of public image and video databases for quality assessment**.** *IEEE Journal* of *Selected* Topics in *Signal* Processing, 6, 616-625.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A Bayesian framework for saliency using natural statistics. *Journal of Vision, 8(7):32,* 1–20.

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision, 11(3):9,* 1–15.